An overview of the OAIS and Representation Information

Digital Curation Centre – Imperial College Internet Centre Workshop
Imperial College, London
16th October 2007

Manjula Patel
UKOLN, DCC
University of Bath, UK
m.patel@ukoln.ac.uk

Funded by:
JISC
e-Science

Presentation Outline

The OAIS Reference Model

- Background
- Concepts
- Functional Model
- Information Model
- Representation Information and Networks
- Responsibilities and Conformance

Registry/Repository of Representation Information

- DCC Development
- RRoRI
- Case studies: crystallography, engineering

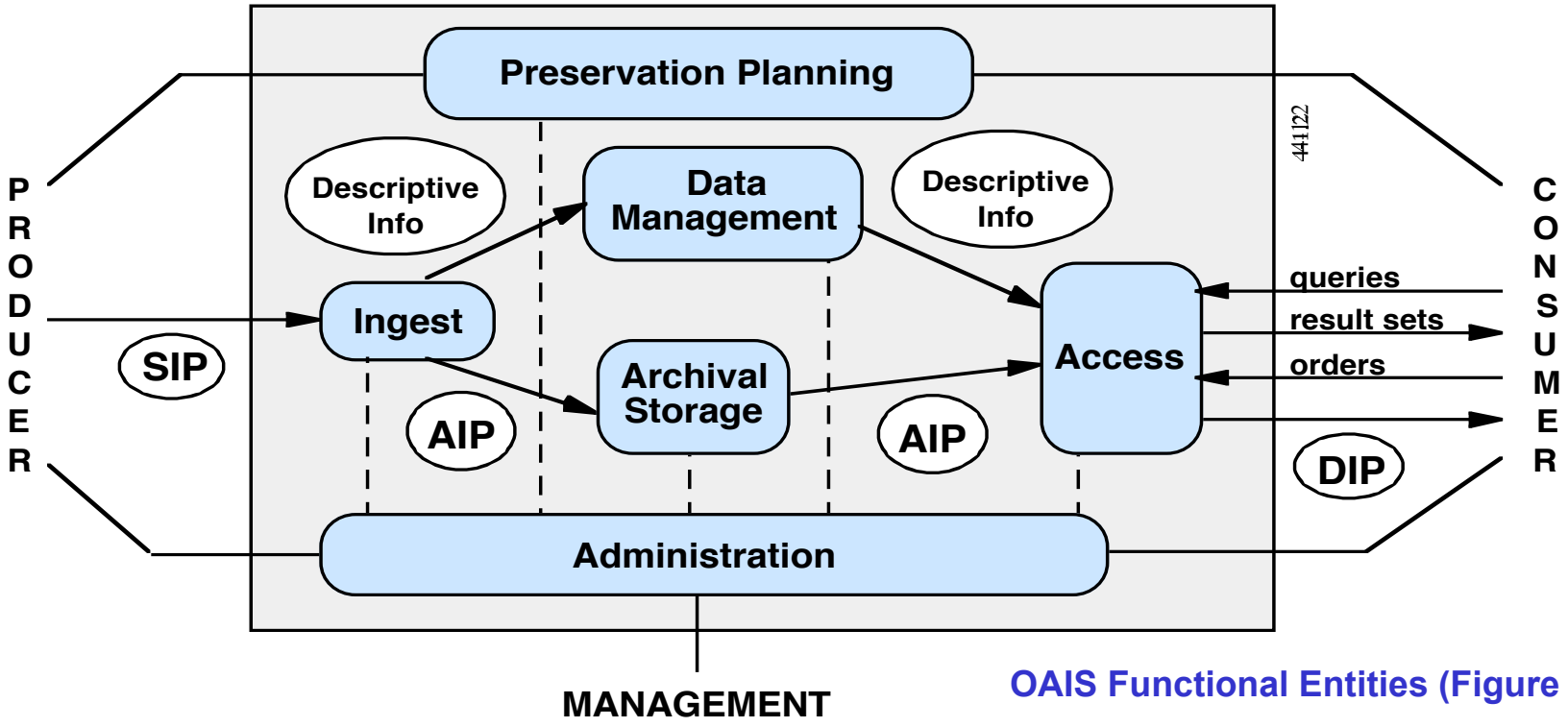
OAIS Background

- OAIS -*Reference Model for an Open Archival Information System* http://www.ccsds.org/documents/650x0b1.pdf
- Development led by the Consultative Committee for Space Data Systems (CCSDS)
- Adopted as ISO 14721:2003 (currently under review)
- "Open" refers to development of the model in an open forum
- Reference Model, not a blueprint for implementation
- Establishes a common framework of terms and concepts
- Identifies the basic functions of an OAIS
- Defines an information model
- Three major areas of influence:
  - Preservation metadata schemas
  - Architecture and system design
  - Conformance criteria for archival repositories

OAIS Definition and Selected Concepts

- OAIS: "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community"
- Designated Community: Community of stakeholders and users that the OAIS serves
- Knowledge Base: A set of information, incorporated by a user or system, that allows that user or system to understand the received information
- Information Object: Data Object + Representation Information
- Representation Information: **any** information required to render, interpret and understand digital data
- Information Package: Content Information + Preservation Description Information + Packaging Information (Submission, Archival and Dissemination Information Packages)
- Preservation Description Information: Provenance, Context, Reference, Fixity information

OAIS Functional Model

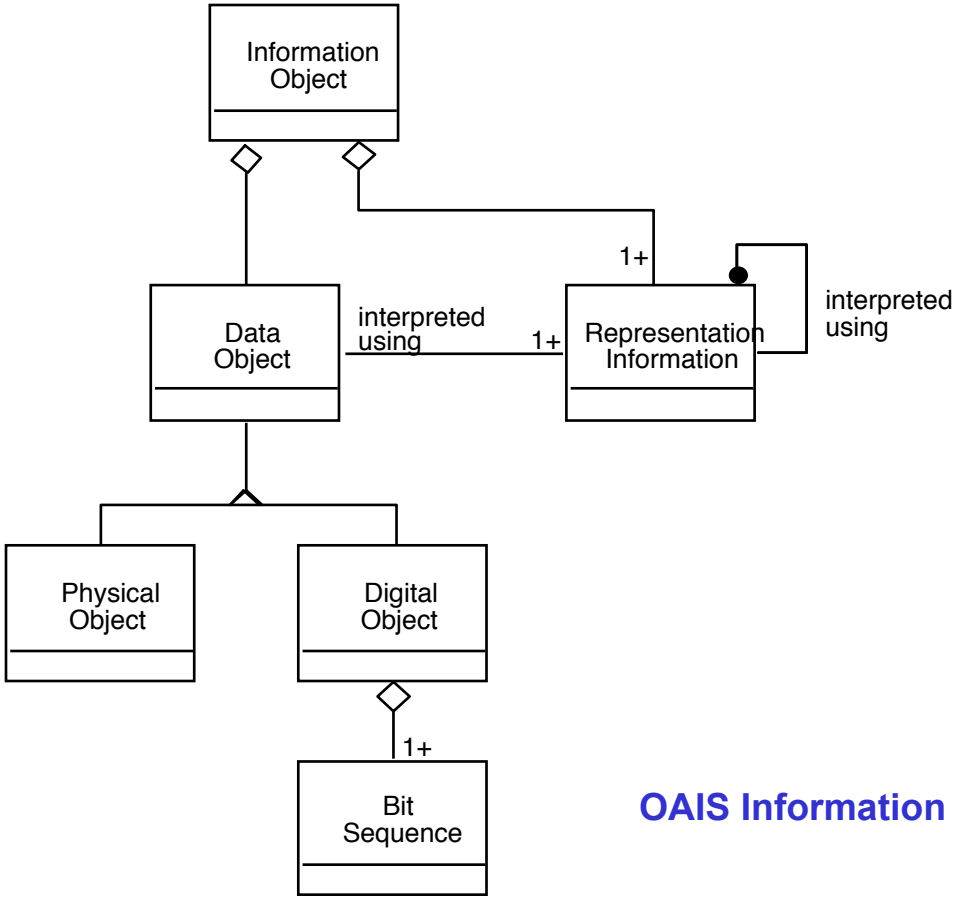


OAIS Functional Entities (Figure 4-1)

OAIS Functional Entities

- Ingest: services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS
- Archival Storage: services and functions used for the storage and retrieval of AIPs
- Data Management: services and functions for populating, maintaining, and accessing a wide variety of information
- Administration: services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis
- Preservation Planning: services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community
- Access: services and functions which make the archival information holdings and related services visible to Consumers

OAIS Information Object

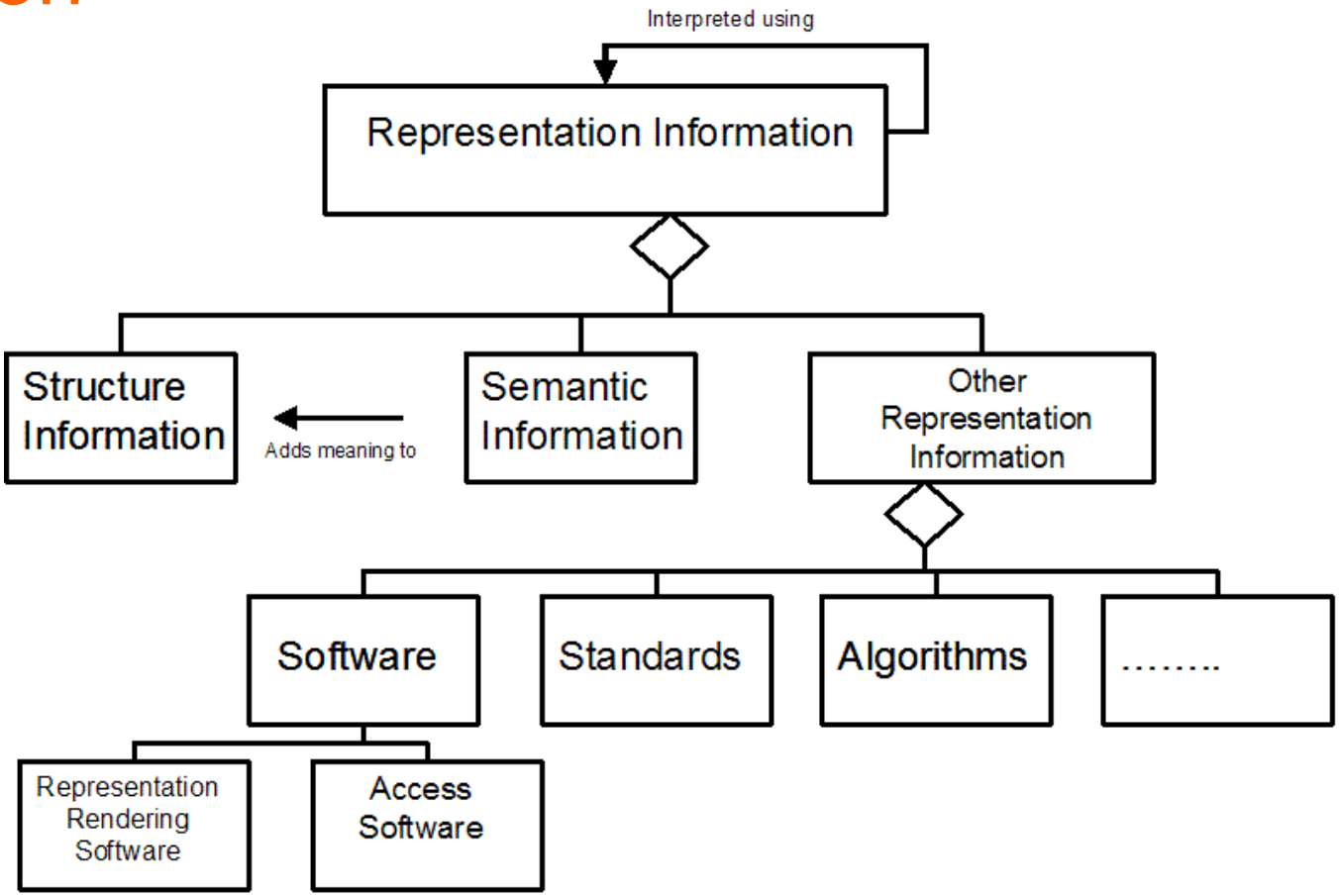


OAIS Information Object (Figure 4-10)

OAIS Representation Information (RI)

- Representation Information: **any** information required to render, interpret and understand digital data (includes file formats, software, algorithms, standards, semantic information etc.)
- Representation Information is recursive in nature
- Essential that Representation Information itself is curated and preserved to maintain access to (render and interpret) digital data

OAIS Representation Information Classification



Types of Representation Information

- Structure

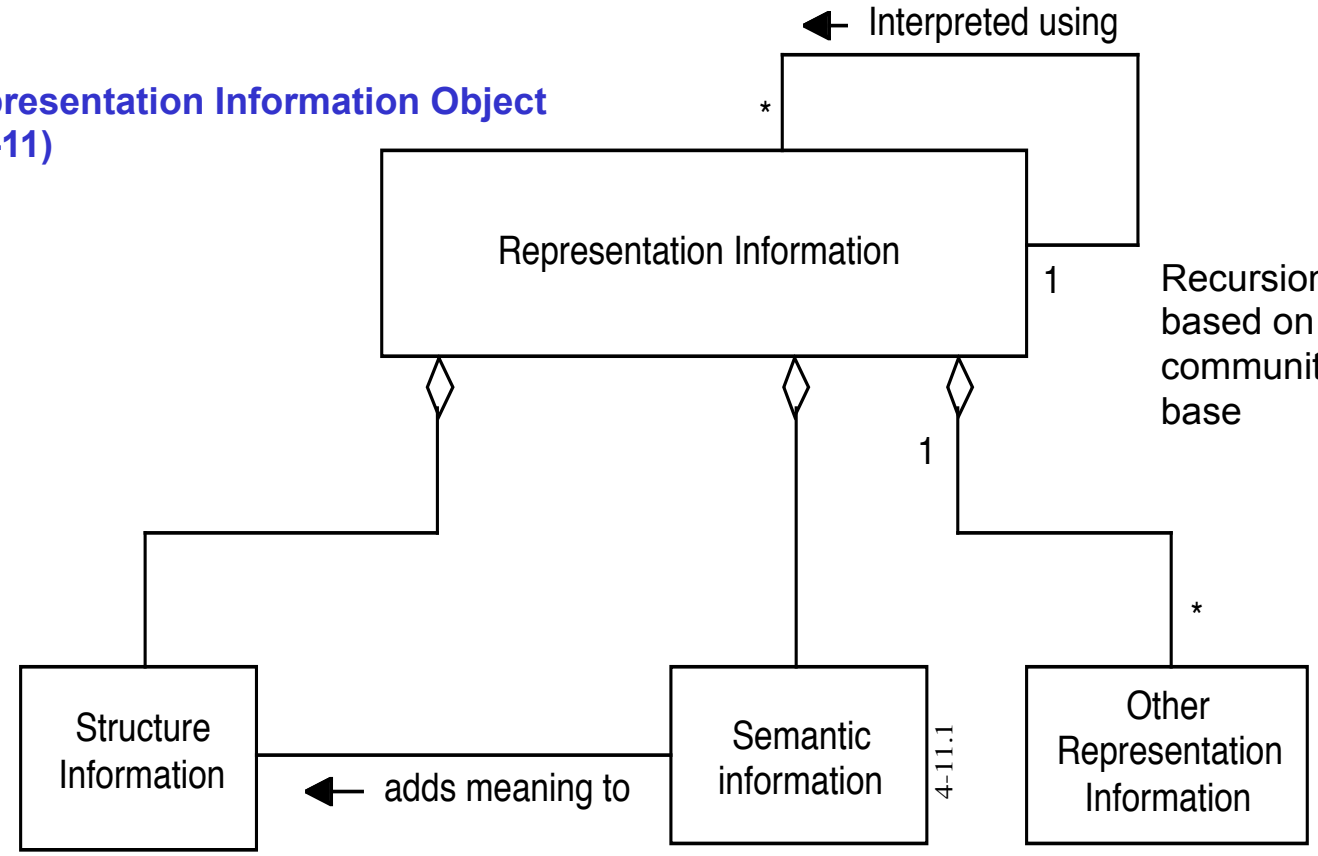
  e.g. file formats for text, images, audio, moving images, datasets, 3D models
- Semantic

  e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri
- Other

  e.g. software, algorithms, standards, time dependent information, actions, processes

OAIS Representation Information Network

OAIS Representation Information Object (Figure 4-11)



OAIS Responsibilities and Conformance

- OAIS Mandatory Responsibilities:
  - Negotiating and accepting information
  - Obtaining sufficient control of the information to ensure long-term preservation
  - Determining the "designated community"
  - Ensuring that information is "independently understandable"
  - Following documented policies and procedures
  - Making the preserved information available

- Many repositories or preservation tools claim OAIS compliance:
  - e.g., DSpace, OCLC Digital Archive, METS, LOCKSS etc.

OAIS…More

- Conformance and Certification
  - OCLC/RLG Digital Archive Attributes Working Group (Report on Trusted Digital repositories, 2002)
  - RLG-NARA Task Force on Digital Repository Certification (Draft checklist for self-certification, August 2005)
  - Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist (CRL, Feb. 2007)
- Archival Information Units and Archival Information Collections
- Information Package transformations, e.g. for Ingest and Access
- Preservation perspectives:
  - Migration e.g refreshment, replication, repackaging, transformation
  - Preservation of look and feel (e.g. emulation, virtual machines)
- Archive interoperability, e.g. P2P, federation

DCC: Development

- Led by David Giaretta, Science and Technology Facilities Council
- "DCC Approach to Digital Curation" sets out the path for development activities based on the OAIS http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCApproachToCuration
- Monitoring international standards
- Development of a Registry/Repository of Representation Information (RRoRI)
- Recommendations for tools and methods for generating Representation Information
- Creating test-beds for digital curation tools
- Creating auditing and certification processes for trusted repositories

RRoRI

- Representation Information is the key to long-term access
- RRoRI should be OAIS compliant
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI which other archives can rely on
- Investigate how RI fits into the work of other projects and initiatives
- Work now being undertaken jointly with the CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
  - Integrated Project co-funded by EU FP6 Programme, April 2006

RRoRI: Curation Persistent Identifier

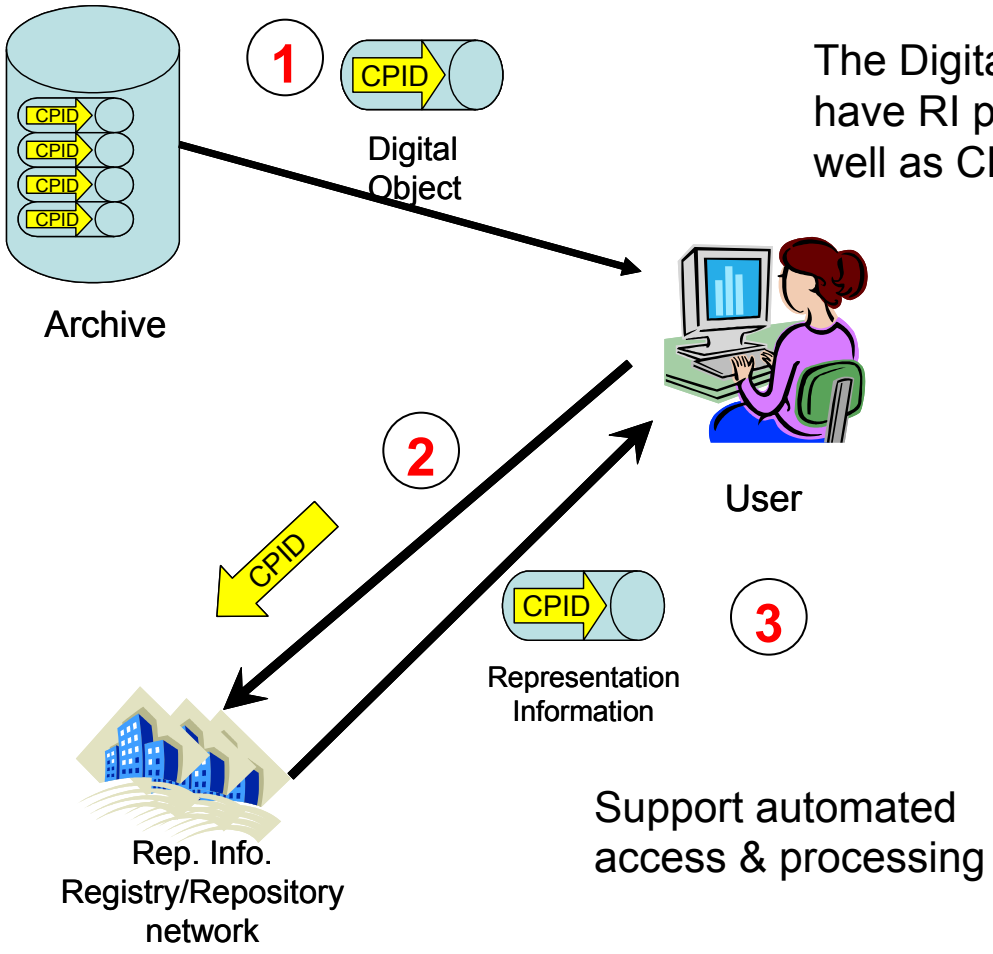
- Idea of RI is the key
  - Information Object: a specific object to be archived
  - RI: all information required to interpret and render the object
  - RI Label: used to connect RI to an Information Object
- RI label serves as a mechanism for accessing RI in the RRoRI
  - A label attached to each digital object
  - Label should identify RI
  - Provides mechanism for combining individual RI components
  - May be a structured digital object itself (to cope with packaging of multiple objects)
- RI label has a Curation Persistent Identifier (CPID)

Use of CPID

1 User gets data from archive. Data has associated Curation Persistent Identifier (CPID)

2 User unfamiliar with data so requests RI using CPID

3 User receives RI – which has its own CPID in case it is not immediately usable



The Digital Object could have RI packed with it, as well as CPID

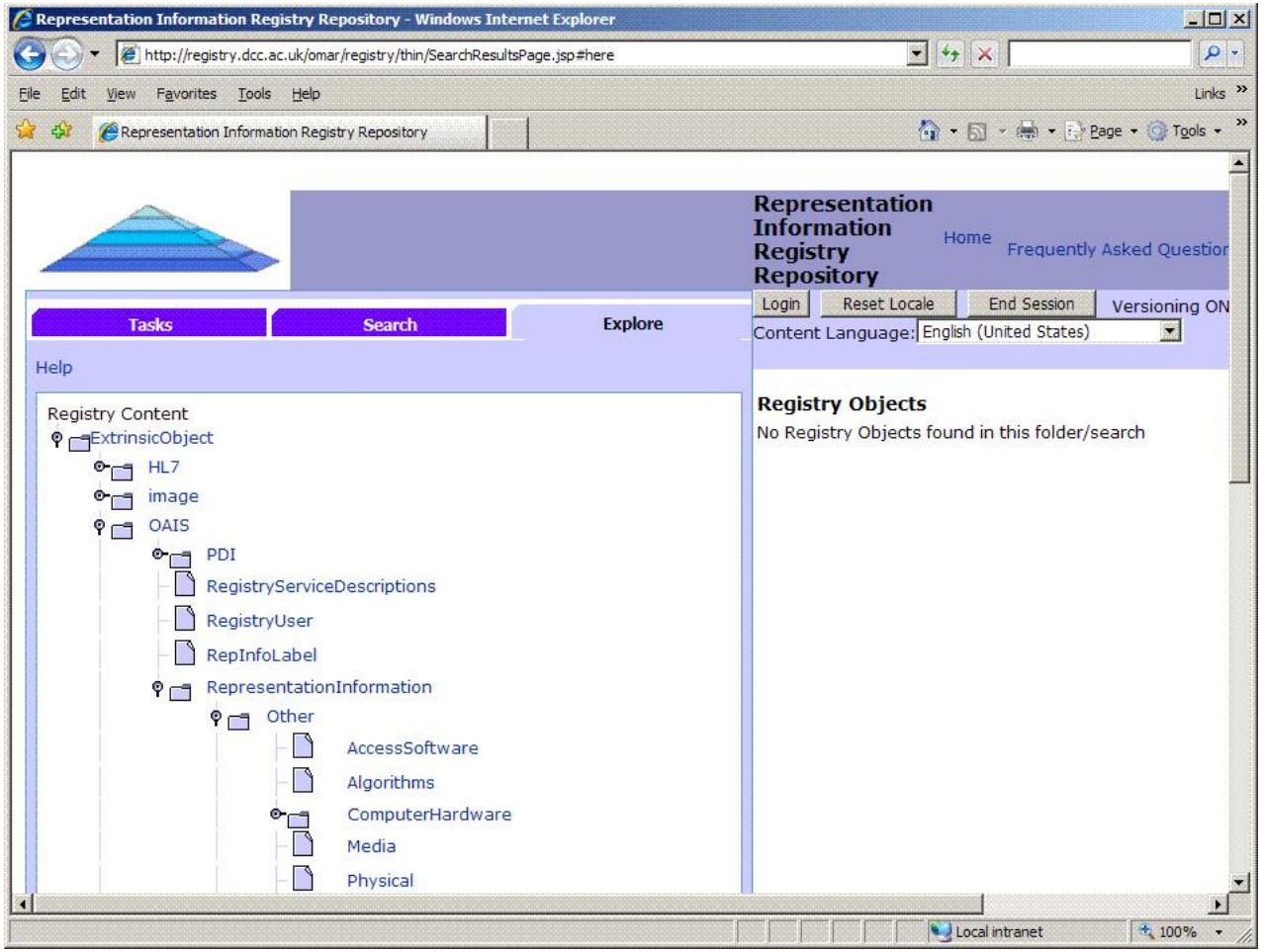
Support automated access & processing

David Giaretta, 2007

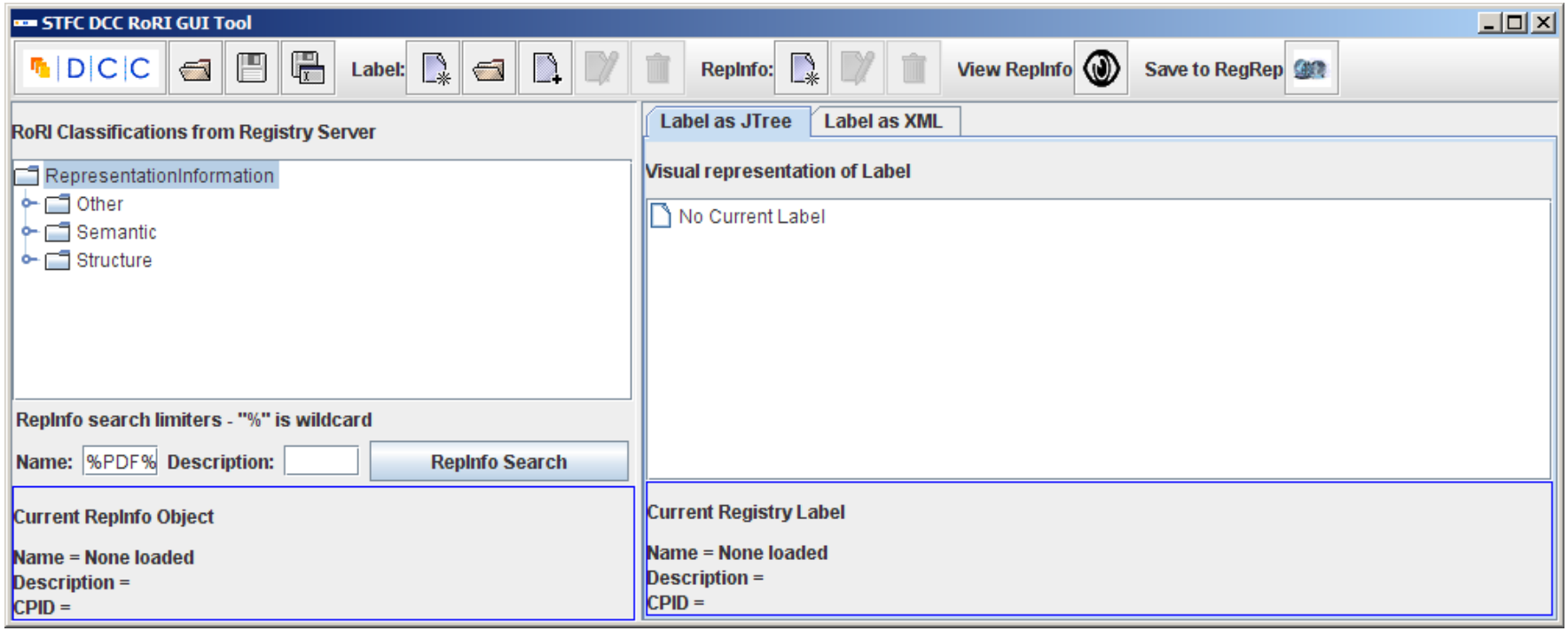
RRoRI: Technical Platform

- freebXML registry
- SOAP messaging
- Java API
- HTTP access
- GUI Tool (label creation and RI ingest)

RRoRI Web access



GUI Tool



Facilitates creation of RI labels and ingest of RI

Two case studies (preliminary work)

- eBank-UK Phase 3 study
  - JISC-funded from Sept 2006-June 2007
  - UKOLN (lead), University of Southampton (NCS), University of Manchester
  - Open access to datasets
  - Linking research data to publications and scholarly communication

- Knowledge & Information Management through life (KIM-GC)
  - 8 Academic partners
  - Industrial partners: construction; aerospace, defence suppliers; MOD; NHS
  - £5.5 million total funding, £3.68 million EPSRC/ESRC, Oct 2005-Oct 2008
  - Develop tools and techniques for sustainable representation of product, process and design rationale
  - Develop approaches to learning about products in service – the performance of the artefact and its impact on users
  - Investigate the dynamics of knowledge use throughout the life-cycle of complex product-service systems, and make recommendations for improved effectiveness
  - Develop an intellectual framework for the above

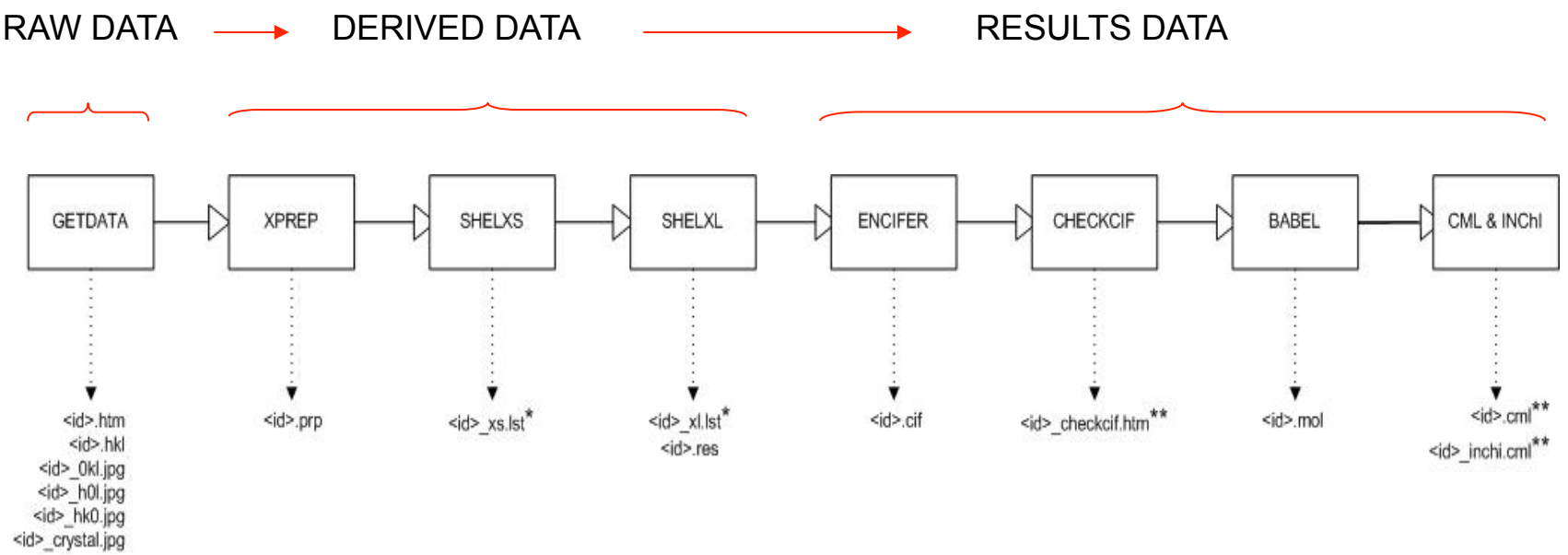
eBank-UK Study

M. Patel and S. Coles, "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation", Sept. 2007

http://www.ukoln.ac.uk/projects/ebank-uk/curation/

- audit and certification (TRAC, DRAMBORA, NESTOR, ISO International repository audit and certification BOF Group)
- OAIS and Representation Information
- eBank-UK application profile and preservation metadata
- e-Prints.org repository platform

Crystallography Workflow



Simon Coles, 2006

- Initialisation: mount new sample, set up data collection
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures
- Refinement: refine structure
- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report

Capturing RI: eCrystals Repository

- Bounded domain (within an academic environment)
- Limited number of stakeholders
  - International Union of Crystallography (IUCr)
  - UK National Crystallography Service (NCS)
  - Cambridge Crystallography Data Centre (CCD)
  - Royal Society of Crystallography
  - Chemistry Central
  - Reciprocal Net
- Open standards and software e.g. checkcif, CML, INChI
- Culture for sharing data
- Well-established workflow for crystallography experiments
- One dominant file format (CIF) - international exchange format
- http://homes.ukoln.ac.uk/~lismp/IDCC2007/RINetCIF.htm

Capturing RI: KIM-GC Project

- Engineering is a broad area (mechanical, electrical, civil; architecture, construction, defence etc.)
- Vested commercial interests
- Proliferation of proprietary file formats
- Closed software solutions
- IGES 5.3: first popular exchange format (STEP still immature)
- http://homes.ukoln.ac.uk/~lismp/IDCC2007/iges.html

Conclusions

- Need digital curation throughout the useful lifetime of digital data
  - Maximise potential of digital data
  - Maximise investment in digital data
  - Curation should be planned for from the outset
- A preservation strategy based on RI depends on a global, well-engineered, distributed network of RI
  - Needs coordination and collaboration on a global scale
- Domain expertise required for creation of comprehensive RI networks
- Actual task of creating RI networks is time-consuming and non-trivial
  - Need simple and automated tools and procedures
- Likely to be gaps in global networks of RI
  - Business case for using a store of RI is clear, however the case for submitting RI to the global effort is less clear

Selected References

- OAIS Reference Model: http://www.ccsds.org/documents/650x0b1.pdf
- DPC Technology Watch Report on OAIS model by Brian Lavoie (OCLC Research): http://www.dpconline.org/
- Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist (CRL): http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91
- RLG/NARA Task Force on Digital Repository Certification: http://www.rlg.org/
- DRAMBORA -Digital Repository Audit Method Based on Risk Assessment, March 2007, Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), http://www.repositoryaudit.eu/
- DCC Development White Paper “DCC Approach to Digital Curation under Development”: http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCApproachToCuration
- CASPAR Project: http://www.casparpreserves.eu
- M. Patel and S. Coles, "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation", Sept. 2007 http://www.ukoln.ac.uk/projects/ebank-uk/curation/
- eBank-UK Project http://www.ukoln.ac.uk/projects/ebank-uk/
- Knowledge & Information Management through Life: A Grand Challenge Project http://www-edc.eng.cam.ac.uk/kim/

Questions?

Thank you for your attention

Manjula Patel
UKOLN, DCC
University of Bath, UK
m.patel@ukoln.ac.uk

http://www.dcc.ac.uk/
http://www.ukoln.ac.uk/